DOCUMENT RESUME

ED 137 391 TM 006 197

AUTHOR	Hall, Eugene R.; And Others
TITLE	Training Effectiveness Assessment: Volume II, Problems, Concepts, and Evaluation Alternatives. Final Report.
INSTITUTION	Naval Training Equipment Center, Orlando, Fla. Training Analysis and Evaluation Group.
REPORT NO	TAEG-R-39
PUB DATE	Dec 76
NOTE	53p.; For Volume I, see TM 006 196
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS	Data Collection; *Evaluation; *Evaluation Methods; *Evaluation Needs; Measurement Techniques; Military Service; *Military Training; Problems; Program Effectiveness
IDENTIFIERS	*Navy

ABSTRACT

A study was conducted to clarify issues and problems involved in the assessment of the effectiveness of military training and to evaluate and recommend more objective procedures for determining the effectiveness of Navy training. The study results are reported in two volumes. This volume examines specific problems affecting Navy training evaluation programs. It provides discussions of technical considerations relevant to the conduct of evaluation and training effectiveness assessment. General procedures for assessing the effectiveness of Navy training courses are given and a number of methodological options for evaluation data gathering are described and evaluated. Recommendations are made for improving training evaluation practice and for the establishment of a Training Effectiveness Assessment Center to assist in the planning and conduct of Navy training evaluations. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED137391

TAEG

TRAINING ANALYSIS AND EVALUATION GROUP

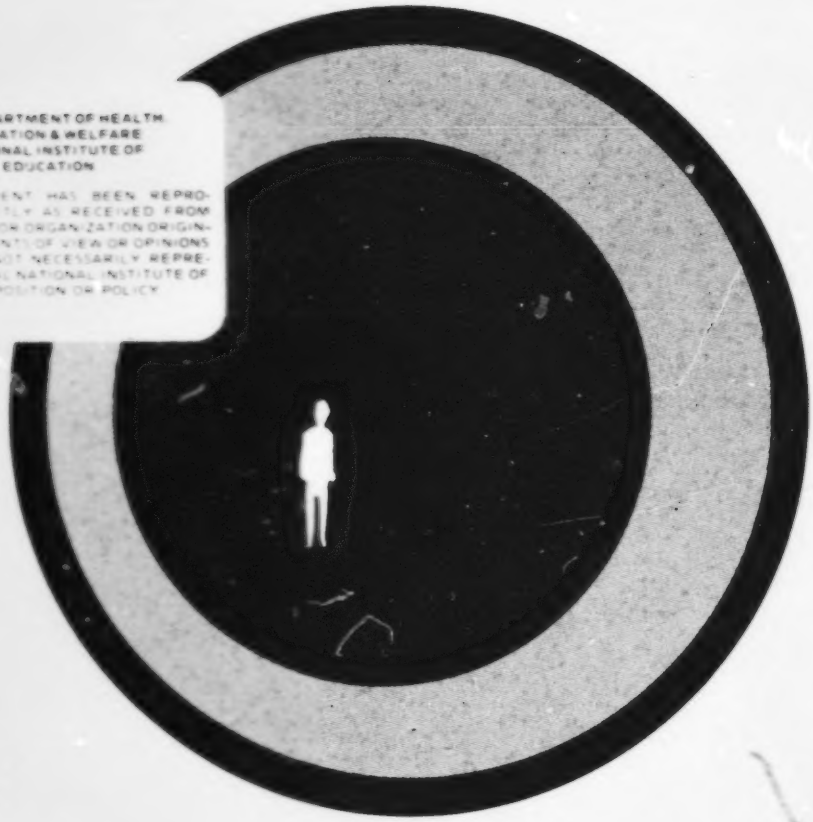
TAEG REPORT NO. 39

TRAINING EFFECTIVENESS ASSESSMENT:

VOLUME II, PROBLEMS, CONCEPTS, AND EVALUATION ALTERNATIVES

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.



FOCUS ON THE TRAINED MAN

TM006 197

2

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION IS UNLIMITED.

DECEMBER 1976

TRAINING ANALYSIS AND EVALUATION GROUP
ORLANDO, FLORIDA 32813

CHIEF OF NAVAL EDUCATION AND TRAINING
NAVAL AIR STATION
PENSACOLA, FLORIDA 32508

Code N-5
5 April 1977

From: Chief of Naval Education and Training
To: Distribution List

Subj: TAEG Report #39: TRAINING EFFECTIVENESS ASSESSMENT:
Volume I, Current Military Training Evaluation Programs
Volume II, Problems, Concepts and Evaluation Alternatives

1. This study, one of a continuing series designed for improved effectiveness in naval training, was conducted to clarify the issues and problems involved in the assessment of the effectiveness of military training and to evaluate and recommend objective procedures for determining the effectiveness of navy training. The measurement of the effectiveness of naval training through the evaluation of trainee performance has been and continues to be a matter of major concern to the Chief of Naval Education and Training. Despite this concern, meaningful product appraisal continues to be less than satisfactory. The CNET has recently emphasized the importance of training appraisal through the establishment of a special division directed by an Assistant Chief of Staff dedicated primarily to this task. A pilot program is now in progress in the Fleet utilizing the Fleet Feedback Data Collection Group as a medium for the collection of evaluative feedback data. This and other initiatives will close the current gap in the navy training system.

2. TAEG Report #39 is authorized for dissemination for the attached distribution list for information purposes.

WORTH SCANLAND, PhD
Assistant Chief of Staff
Research and Program Development

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TAEG Report No. 39	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) TRAINING EFFECTIVENESS ASSESSMENT: VOLUME II, PROBLEMS, CONCEPTS, AND EVALUATION ALTERNATIVES		5. TYPE OF REPORT & PERIOD COVERED Final Report May 1975 - Nov 1976
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Eugene R. Hall, William C. Rankin, Ph.D., and James A. Aagard, Ph.D.		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Training Analysis and Evaluation Group Orlando, FL 32813		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE December 1976
		13. NUMBER OF PAGES 45
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

See also Training Effectiveness Assessment: Volume I, Current Military Training Evaluation Programs

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Assessment
Training
Evaluation
Training Effectiveness
U.S. Navy
Methodology

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A study was conducted to clarify issues and problems involved in the assessment of the effectiveness of military training and to evaluate and recommend more objective procedures for determining the effectiveness of Navy training. The study results are reported in two volumes. Volume 1, Current Military Training Evaluation Programs, describes and assesses current training evaluation programs of the U.S. Air Force, Navy, Marine Corps and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Army. The present volume (II) examines specific problems affecting Navy training evaluation programs. It provides discussions of technical considerations relevant to the conduct of evaluation and training effectiveness assessment. General procedures for assessing the effectiveness of Navy training courses are given and a number of methodological options for evaluation data gathering are described and evaluated. Recommendations are made for improving training evaluation practice and for the establishment of a Training Effectiveness Assessment Center to assist in the planning and conduct of Navy training evaluations.

S/N 0102- LF- 014- 6601

TAEG Report No. 39

TRAINING EFFECTIVENESS ASSESSMENT: VOLUME II, PROBLEMS, CONCEPTS, AND EVALUATION ALTERNATIVES

Eugene R. Hall
William C. Rankin, Ph.D.
James A. Aagard, Ph.D.

Training Analysis and Evaluation Group

December 1976

GOVERNMENT RIGHTS IN DATA STATEMENT

Reproduction of this publication in whole or in part is permitted for any purpose of the United States Government.

ALFRED F. SMODE, Ph.D., Director,
Training Analysis and Evaluation Group

WORTH SCANLAND, Ph.D.
Assistant Chief of Staff
Research and Program Development
Chief of Naval Education and Training

TABLE OF CONTENTS

Section		Page
I	INTRODUCTION	5
	Purpose	5
	Approach	6
	Organization of the Report	6
II	FACTORS AFFECTING TRAINING EFFECTIVENESS ASSESSMENT	7
	Attitudes Toward Evaluation	7
	Administrative Factors	7
	Responsibility for Evaluation	7
	Command Emphasis and Support	8
	Personnel Factors	9
	Manpower	9
	Training	9
III	CONCEPTS IN EVALUATION AND EFFECTIVENESS ASSESSMENT	11
	Evaluation and Training Effectiveness Assessment	11
	Training Goals	12
	Purpose of Evaluation	12
	Evaluation of Course Components	14
	Training Input Variables	15
	Training Process Variables	15
	Requirements for Training Effectiveness Assessment	16
	Choosing Measures of Behavior	16
	Maintaining the Goal Achievement Orientation	17
	Timing for Assessment	18
	Obtaining Evaluation Data	18
	Preparation of Evaluation Plans	18

Section		Page
IV	EVALUATION DATA GATHERING OPTIONS	21
	Attributes of Options	21
	Technical Attributes	21
	Reliability and Validity	22
	Reliability	22
	Validity	22
	Obtrusiveness of Options	22
	Practical Attributes	23
	Development Costs	23
	Administration Costs	23
	Scoring and Analysis Costs	24
	Options	24
	Description of the Options	24
	Objective Data Gathering Procedures	25
	Paper and Pencil Achievement Tests	25
	Work Sample Proficiency Tests	27
	Subjective Data Gathering Procedures	28
	Questionnaires	28
	Interviews	29
	Records and Reports	30
	Variations of EDGO	31
	EDGO Selection	31
	Logical Considerations	31
	Examples of EDGO Selection	35
	Interpretation and Use of Evaluation Data	36

TABLE OF CONTENTS (continued)

Section		Page
V	CONCLUSIONS AND RECOMMENDATIONS.	39
	Conclusions.	39
	Recommendations.	40
REFERENCES		42
BIBLIOGRAPHY		43

LIST OF ILLUSTRATIONS

Figure		Page
1	EDGO Selection Rationale	32

SECTION I

INTRODUCTION

The degree to which the Navy training system meets its established goals affects Fleet readiness and greatly influences the costs of training. Thus, it is important that training be optimally effective. To achieve this goal, knowledge of baseline training effectiveness is needed. It is needed both to determine the value of current training and to provide objective bases for controlling the quality of the training system.

Unfortunately, definitive information about the effectiveness of Navy training is not routinely available. At present, training effectiveness is determined largely by rational assessment and the intuitions of personnel intimately involved in the training process. Information obtained in this way tends to be biased and the detail necessary for improving specific aspects of training is rarely provided. More objective, systematic means are needed for determining training effectiveness and for obtaining information suitable for training quality control.

Recognizing this need for definitive training effectiveness information, the Chief of Naval Education and Training (CNET) tasked the Training Analysis and Evaluation Group (TAEG) to develop an assessment capability for determining the effectiveness of Navy training. Emphasis was to be on the identification and development of means for conducting such evaluations.

PURPOSE

The overall study was concerned with organizing information relevant to the assessment of training effectiveness within a military setting. It was also concerned with the identification and evaluation of methods for assessing training effectiveness. This study is a prelude to the subsequent development of standardized assessment procedures which can be applied on a programmatic basis within the Navy.

The study was conducted in two parts. Part I was concerned with a review of current military training evaluation programs. It is reported separately as Volume I of this report. Part II of the study is reported here. The specific objectives of this portion of the study were to:

- Identify and evaluate factors which affect the establishment and conduct of training effectiveness assessment (TEA) efforts within the Navy
- Clarify and provide technical background information regarding training evaluation concepts and procedures

- Examine and evaluate various methodological approaches for obtaining data for TEA
- Develop recommendations for Navy conduct of TEA.

APPROACH

Information concerning the problems of conducting training effectiveness assessment programs within the Navy was obtained through review of current Navy efforts in this area and by discussion with operational and research personnel involved in aspects of training assessment. A limited literature review was conducted to obtain information about evaluation concepts and procedures. Literature sources were also used to identify various methodological approaches suitable for assessing training effectiveness. Recommendations for Navy conduct of training effectiveness assessments were developed by extraction and synthesis of information from the sources above.

ORGANIZATION OF THE REPORT

The remainder of the report is organized into four sections. Section II presents a discussion of various problems which affect training effectiveness assessment efforts within the Navy. Solutions to these problems are needed for the establishment and conduct of meaningful training appraisal programs. Evaluation concepts and requirements for evaluation plans are presented in section III. Section IV describes a number of methodological options for obtaining evaluation data and provides decision rules and rationales for selection of options for given cases. Conclusions and recommendations for training effectiveness assessment within the Navy are presented in the final section of the report. A special bibliography provides titles of works where more detailed information may be obtained on particular evaluation data gathering options.

SECTION II

FACTORS AFFECTING TRAINING EFFECTIVENESS ASSESSMENT

A number of factors adversely affect current Navy training evaluation programs. Attitudes toward evaluation, administrative provisions for its accomplishment, and certain personnel considerations limit the value of these programs. As such, they represent problems to be overcome if more systematic and objective means for assessing effectiveness are to be developed, accepted, and used by the Navy.

ATTITUDES TOWARD EVALUATION

A fundamental problem for the development and acceptance by the Navy of TEA programs concerns the attitudes of Naval personnel toward evaluation. Currently, these attitudes reflect less than enthusiastic support for evaluation. They range from totally negative through indifferent to lukewarm. The reasons are many--and complex. Some are subjective, and others have more objective bases.

On the subjective side, it is probable that at least some training managers are sensitive to the threats implicit in evaluation. The perception is that evaluation results could be used to affix blame for training program deficiencies. Thus, the positive potential of evaluation for training program improvement may not be realizable. Instead a lack of interest in, or even opposition to, evaluation may arise. On more objective grounds, it is also probable that managers who might otherwise favor evaluation to improve training do not have the necessary resources (e.g., trained personnel, time, money) to implement and conduct meaningful programs. Hence, they feel that they can ill afford to divert scarce resources to programs that have had marginal impact.

ADMINISTRATIVE FACTORS

Certain administrative factors contribute to the creation and/or reinforcement of existing attitudes toward evaluation. These administrative factors have important implications for the quality and value of evaluation programs. They include assignment of responsibilities as well as command emphasis and support for evaluation.

RESPONSIBILITY FOR EVALUATION. Currently, responsibility for maintenance of the quality of training through the use of evaluation mechanisms rests ultimately at the training course level. Except for unusual circumstances, higher levels of command(s) are not directly involved in the process.

The CNET has ultimate responsibility for the training system, but, appropriately, much of this has been delegated to lower functional

levels within the Naval Education and Training Command (NAVEDTRACOM). Organizational functions, established at various command levels (e.g., within CNET, CNTECHTRA, CNET Support), are assigned limited evaluation responsibilities. By policy, these groups function largely in evaluation support roles rather than in controlling or directing roles. They provide general guidance, usually in the form of instructions or tools (e.g., checklists), for use by the individuals (usually instructors) who design and conduct particular courses. It appears that training evaluation, when conducted, is largely at the initiative of local training personnel who are responsible for it. This responsibility includes obtaining feedback data from the Fleet for use in course improvement. Information concerning if or how well evaluation is accomplished, or how results are used for training improvement, usually is not disseminated above the course level. It is not known if corrective action mechanisms for removing training deficiencies exist at local course levels.

COMMAND EMPHASIS AND SUPPORT. Despite the existence of command directives and instructions regarding training evalution, command emphasis on, and support for, evaluation are less than desirable. Evaluation of training has not been given the command attention that it merits. Enforcement mechanisms for insuring compliance with directives are not clearly apparent. Command organizations assigned evaluation responsibilities seem to function in passive monitoring roles and routine accounting of course status is apparently not required. Within some training organizations, there is an apparent lack of interest in assessing the effectiveness of training. Comments such as "we don't have time" or "we are interested only in production" are not uncommon. The belief that "things are good enough" and the lack of an overall firm command emphasis on and commitment to conducting evaluations of training undoubtedly contribute to the apparent lack of interest.

Command support for evaluation seems, most often, to take a form of "noninterference" in training matters. The policy makers seem to prefer to leave the issue to lower levels. At command levels, a prevalent view has been that the instructional staff should define training needs, construct training programs, and deliver and evaluate instruction. The assumption is that the instructional staff, being technically competent in the subject matter at hand, is best qualified to define training needs and accomplish necessary instructional and evaluation functions on a not-to-be-interfered-with basis. Instructions and directives written at administrative levels which pertain to training appraisal provide only minimal guidance. Generally, this concerns only the type of activities required for training appraisal. Specific guidance in how to conduct such appraisals, expert assistance, and manpower, for their prosecution is not provided.

PERSONNEL FACTORS

Personnel inadequacies also limit the value of current training evaluation programs. The availability of adequate numbers of personnel to conduct training evaluations (i.e., raw manpower) and their training for these roles are at issue here.

MANPOWER. On the manpower side, when additional requirements are imposed upon training organizations, additional personnel to accomplish the new functions are not often provided. CNET Instruction 1540.6 (Establishment of Curriculum and Instructional Standards (CIS) Offices/Departments), for example, levies requirements for many specific evaluation activities. But, the Instruction specifically states that additional personnel will not be assigned to fulfill the functions required by it. Consequently, (new) functions, such as evaluation, may either not get accomplished at all, or, at best, may become collateral duties of an already fully-committed instructional staff. The availability of time to perform evaluation functions is a limiting factor in their accomplishment. It probably also engenders a lack of wholehearted interest in such accomplishment.

The practice of assigning evaluation functions to the instructional staff is not desirable from another standpoint; i.e., that of the lack of objectivity of obtained results. Personnel intimately involved in any activity are rarely qualified to evaluate their own efforts objectively. Personal pride in achievement and the belief that one is already doing his best militate against any desire, and perhaps even preclude the ability, to evaluate critically and objectively one's own work. The training evaluation function should be separate from the training process itself to avoid subjectivity and bias. Otherwise, a clear interpretation of evaluation results will not likely be possible.

TRAINING. Lack of adequate training (and experience) for evaluation personnel also affects the value of evaluation programs. Typically, military instructional personnel do not have an evaluation background. Most, probably, do not fully understand the need for, purposes and procedures of evaluation. (These topics are discussed in sections III and IV of this report.) In the past, solutions to the problem of lack of relevant training have been sought by attempting to produce or otherwise obtain easy-to-follow, readily comprehensive guides for "anyone's" use. But, detailed procedural manuals or handbooks for conducting evaluations will probably not be effective substitutes for evaluation training or skilled professional personnel. A study for the United States Army (Ricketson, et al., 1970), investigated the value of a detailed manual, containing step-by-step procedures, for developing training programs (using the Systems Engineering of Training Approach). This study found that the average Army officer was unable to complete the course design process satisfactorily. This was attributed to a lack of in-depth

understanding of specific requirements and a general lack of the necessary technical knowledge needed for correct decision making.

In the field of evaluation many complex decisions are required. Consider, for example, just one of the steps to be taken in conducting internal evaluations of training courses. The evaluator is told to insure that the learning objectives are based on a task analysis. On the surface, this is a simple task requiring only an inquiry. But, it is not enough simply to determine that the learning objectives are based on a task analysis. Good evaluation practice requires that the task analysis data base also be examined to determine if it contains valid information accurately describing, at the proper level of detail, the operations required to perform a particular job. Similarly, determining that stated learning objectives accurately reflect the tasks to be performed on the job is a difficult undertaking. Effective, meaningful and worthwhile evaluation programs also involve complex decision making and require careful attention to details. The proper development of instruments and procedures for obtaining feedback from the Fleet regarding course graduate job performance requires considerable technical knowledge. Interpreting feedback information and correctly using it to modify training courses also requires greater training than is now routinely given to personnel assigned evaluation functions.

SECTION III

CONCEPTS IN EVALUATION AND EFFECTIVENESS ASSESSMENT

There is an apparent lack of clear understanding of the goals, content, and methodology of training evaluation. In addition to the problems described in the preceding section, this also severely limits realization of the full potential of training evaluation programs. Much of this inadequate knowledge undoubtedly stems from the lack of a solid base of literature and practice regarding training effectiveness determinations in a military setting. At present, the primary source of information about evaluation concepts and practice is the voluminous literature of educational evaluation (see, for example, Popham, 1975). Unfortunately, the direct application of this information to the military environment is extremely difficult. By and large, this literature stresses the need for evaluation and discusses the many problems that plague evaluators within conventional educational contexts. Relevant information and practical methods for accomplishing evaluations within a military setting are not directly emphasized.

Information relevant to the evaluation of military technical training is presented in this section. Concepts pertinent to evaluation in general are discussed. Technical considerations for planning and conducting training assessments are also presented. No attempt is made to provide a complete prescription for determining training effectiveness. The discussions rather are intended:

- To clarify evaluation issues which affect determinations of training effectiveness, and
- To familiarize training managers and potential training evaluators with the general requirements and procedures for making such assessments.

EVALUATION AND TRAINING EFFECTIVENESS ASSESSMENT

Within the military, there is a tendency to equate the general idea of evaluation with the more specific notion of training effectiveness determination. Thus, the results of any type of evaluation may be presented as evidence of the training effectiveness of a course. In our view, this is incorrect. Determining the training effectiveness of a course represents a particular type of evaluation. Evaluation, as a generic term, connotes the general theme of determining the worth, quality, or value of something by comparing it to a standard. These standards may be held implicitly or they may be objectively stated. Training effectiveness assessment is concerned with specific information about trainee achievement. It refers to the effect(s) that training has on the students receiving it. The desirability of these effects

is "evaluated" by reference to the goals of the training course. Evaluation, in the general sense, is concerned more directly with the quality of training than with its effects. Thus, in this report, the term "training effectiveness" is used exclusively to refer to (measures of) the degree to which a training course or system achieves the goals established for it. The term "evaluation" is used in its more general meaning. Obviously, the quality of training greatly influences the effects that will be produced by it. Consequently, continuous evaluation is required to insure that training continues to produce desired results. A first problem, however, is to determine what these results are.

TRAINING GOALS. The general goal of military technical training is to change human behavior in desirable ways. The "changes" desired in behavior are contained in specific course goals. Thus, measures which reflect student achievement with respect to the goals of training are measures of training effectiveness. General statements of course quality are not. To determine training effectiveness, the training goals must be translated into behavioral terms so that trainee achievement can be measured and evaluated against the course goals.

Two sets of goals may exist for a given training course: the end-of-course objectives established by the training unit (or the instructor) and the requirements of the job for which the training was given. Job performance ultimately is the final test of training value. Hopefully, the end-of-course objectives will validly reflect these job requirements. Achieving the course objectives then will assure that the student can, in fact, perform the job for which the training was given. Since, frequently, there is less than a perfect correspondence between the course objectives and the job requirements, it is necessary to validate the objectives, and training, in the operational context. Here, an "external evaluation" (i.e., determining by some appropriate technique that course graduates can, in fact, fulfill job requirements) is needed to determine the ultimate value of training. Courses may be considered effective, however, if they produce graduates with skills and knowledges consonant with either of the "two" sets of goals. However, the evaluator should be clearly aware that while both goals will provide an assessment of course effectiveness, they are not strictly equivalent in terms of the validity of training. Job performance is both a measure of effectiveness and validity. Achievement of course objectives is a measure of effectiveness and <u>may</u> be a measure of validity of training to the extent that it is correlated with performance.

PURPOSE OF EVALUATION

The purpose for which an evaluation is (to be) conducted is all important in deciding how to proceed. This applies to evaluation in general as well as to assessing trainee goal achievement.

Evaluation has both a general purpose, or goal, and, within a given context, it has highly specific goals. At a general level, the purpose of evaluation is to obtain information that can be used for decision making about training. The specific goals are determined by the particular information needed. These specific goals should guide evaluators in establishing evaluation plans and conducting evaluations. The kinds of decisions that need to be made may dictate the data necessary to collect, its attributes, and the form in which it is collected.

Different kinds of decisions are necessary at different management levels. At the CNET level, the information necessary for policy making decisions or for resource allocation may differ in very specific ways from that needed at the training unit level. At the CNET level, depending upon specific decision needs, rather gross, summary measures reflecting how well the training organization is functioning (e.g., measures of system effectiveness) are probably sufficient for many decision needs. At the training unit level, very specific information (such as the nature of the mistakes that students make) is necessary for course revisions. Evaluation design must be based on satisfaction of particular information needs. To determine training effectiveness, the information needed is defined as measures of students' achievement of course goals. Procedures can be specified for obtaining this information. But while the job that needs to be done is deceptively simple in concept it is exceedingly difficult to accomplish in practice. (The next section discusses "procedures" for determining training effectiveness and describes many of the difficulties that must be overcome.)

Conceivably, the purpose of evaluation may sometimes be only to determine the effectiveness of a given course. There may be no interest in changing (or no reason to change) a course. But, if it is found that the measured level of training effectiveness is not satisfactory, then the assessment technique should provide information that can be used to improve the training program. Usually, some gross information about training deficiencies will be collected as part of the TEA effort. Generally, this will be of the form: X number of individuals cannot perform Y tasks. While this information is necessary and useful, it is not usually sufficiently detailed to discover what to change in training to correct the observed deficiencies. If training is to be improved, then a second level of analysis is required. A third level of analysis (or evaluation) may also be required to determine how to change it. Note that the purpose of evaluation is now different. Therefore, the data of interest and the procedures to follow for acquiring it will also differ.

At the second level of analysis, it may be necessary to examine different parts of the training program (e.g., course content, teaching methods used, hands-on practice available) to determine the changes needed to achieve greater training effectiveness. This will require an evaluation of "suspect" portions of the course. Internal evaluation

procedures will be required for at least portions of such analyses. Here, an examination and critical assessment of the procedures and content of instruction within the environment in which the training occurs would be indicated. This "type" of evaluation is conducted against criteria of "good" educational practice. Items such as the style and clarity of the training objectives and training content, their relevance to job requirements, availablility of remedial and counseling assistance, quality and quantity of training aids and other media available, etc., should be assessed.

It is important to recognize, however, that internal evaluation does not constitute an assessment of training effectiveness. It is concerned, rather, with the "correctness" of the processes, procedures, and content of instruction. Internal evaluation directly assesses the quality of training--not its effects. If training effects are not what is desired (i.e., the desired effectiveness is not being achieved), then all portions of a course must be examined (evaluated) to determine why.

The third level of evaluation may be necessary when it cannot readily be determined if a proposed change will, in fact, improve overall training effectiveness. In this case, an evaluation of alternative instructional methods may be required before they are recommended for use in a training program. Here, a relative evaluation of training effectiveness would be involved. The question of interest concerns which methods, media, or techniques in relation to others are more (or less) effective in achieving desirable instructional outcomes. While experimental methods may be best for evaluation of differential effectiveness, other less-demanding evaluation techniques may also be suitable depending again on elements of the specific situation.

To summarize briefly, training effectiveness information should be collected to determine the degree of achievement of training goals. An unsatisfactory level of achievement indicates the need for closer examination (and evaluation) of elements of the course to determine possible reasons for the less than desirable effectiveness. Thus, evaluation efforts may be focused on different "parts" of the course in attempts to determine the specific areas needing improvement. These are discussed next.

EVALUATION OF COURSE COMPONENTS

Any training course has three essential components, or elements: an input, a process of instruction, and a product of that instruction. The effectiveness of training can only be determined by evaluation of the training product--the course graduate. The nature of the input and the instructional process determine the final state of that product. If it is determined by evaluation of that product that a course is not

effective, then the other two elements must be carefully examined (evaluated) to discover the changes necessary for improving graduate quality.

TRAINING INPUT VARIABLES. There are many inputs to a training course which may affect training outcomes. Conventionally, inputs to training are thought of as a student population having certain characteristics (e.g., aptitudes, educational levels). Other variables, such as physical and environmental factors (e.g., seating and lighting arrangements, temperature), fiscal support, externally supplied training objectives, however, may also be considered as training input variables. The operation of these input variables and/or changes in them may be more responsible for training failures (i.e., the lack of desirable effects) than the way in which a course is structured or conducted. For example, a sudden or sharp rise in attrition rates, setbacks, or other failures to complete training satisfactorily could be due to changes in the qualification of students entering the course. Similarly, other input variables such as poor lighting could be responsible for unsatisfactory student achievement. Thus, input factors should be evaluated routinely to determine their continuing quality and adequacy and also to determine their contribution to, or effects on, overall training effectiveness. These should be assessed prior to concluding that changes are needed to the course itself. The CNTECHTRA A10 Manual contains checklists for evaluating quality of some input variables. Changes in student characteristics over time, which may affect training outcomes, can be detected by keeping records of student entering capabilities.

TRAINING PROCESS VARIABLES. More often, training failures will be due to defects in the instructional process. Inappropriate content and faulty instructional procedures will affect student goal achievement. Thus, the process of instruction should be carefully evaluated. At present, Navy instructions stipulate that this should normally be accomplished as part of an "internal evaluation." Here, the content of instruction, the media used, the instructional practices, procedures, materials, and strategies should be periodically examined to insure that their design is effective for learning and that they are being used in effective ways. Typically, these types of evaluations are made intuitively ("in the best judgment" of the assigned instructional staff). Sometimes, they are made by comparisons of the characteristics of instructional elements or aspects to lists of desirable characteristics or attributes (see, for example, the CNTECHTRA A10 Manual). They should be, but too frequently are not, conducted with reference to the job requirements for which the course is given. (Note that in evaluating the process of instruction, the real concern is with the differential contribution of the different components of instruction to the total effectiveness of the course; i.e., the training effectiveness of individual components of the course and not the effectiveness of the total course. Here, the methods of evaluation used, the most powerful of which are

experimental methods, are different from those used to determine overall course effectiveness.)

REQUIREMENTS FOR TRAINING EFFECTIVENESS ASSESSMENT

To determine training effectiveness, data reflecting changes in behavior (specifically, changes in skills and knowledges) must be obtained. This information is used to ascertain if, or how well, the course meets the objectives for which it was established. Thus, there are two principal requirements for TEA: (1) defining the behavioral information needed and appropriate evaluation criteria and (2) obtaining that information by some appropriate means.

CHOOSING MEASURES OF BEHAVIOR. As noted previously, measures of student behavior which reflect achievement of course goals are needed to determine a course's effectiveness. Thus, it is necessary that the goals be explicitly stated so that they can be translated into human performance terms. For those courses which have objectives explicitly stated in behavioral terms, this translation has already been made. Thus, student performance at the end of training can be directly compared against the objectives for a ready determination of training effectiveness. Unfortunately, many Navy training courses, at the present time, do not have goals which are stated in explicit behavioral terms. Most often, course goals are stated in such general terms that it is not readily apparent what the course is intended to achieve. In these cases, analytical effort (e.g., job analysis, instructor interviews, study of instructional materials) may be required to identify and clearly state the goals so that student achievement may be compared to them.

Frequently, an evaluator must also develop standards to attach to the behavioral statements that reflect achievement of the training goals. These standards impose limitations, or tolerances, on the behavior to be observed. They specify how well the student must perform (a la formal statements of course training objectives). If his behavior, verbal or motor, is within these tolerance limits, then this is accepted as evidence of achievement of specified skills and knowledges.

If there are no explicitly stated course objectives, and they cannot be developed, no set rules can be given for selecting or identifying the behavior to measure for determining training effectiveness. In such cases, measures of central tendency (i.e., means, medians, modes) could be used. Performance of students on a common achievement test, for example, could be compared to that of certain normative groups. For example, if the mean score of a current class is equivalent to the mean score of previous classes, then satisfactory training effectiveness can be inferred. The validity of the course is, however, uncertain. If the course does have explicit objectives, then the number of these achieved (or time required to achieve) could be used to indicate effectiveness.

Missed items or those requiring longer times to complete would provide diagnostic information about areas where course improvement might be indicated. Percentiles, pretest and posttest scores, or other measures of amount and direction of change in student skills and knowledges could also be used to express training effectiveness. Other measures appropriate for reflecting trainee learning are discussed in section IV which presents various techniques for obtaining evaluation data.

To express overall course effectiveness, the achievements (i.e., scores) of individual course graduates must be appropriately summarized. A second type of standard is then imposed to determine if the course is of acceptable effectiveness. The second type of standard is concerned with the "goodness" of the summary values of the measures of human behavior that are used to represent or reflect training effectiveness. If the evaluator elected to reflect training effectiveness by using a mean score on a common test, then a criterion value must be placed on this mean for it to be accepted as evidence that the course is achieving an overall satisfactory level of effectiveness. Courses which fail to achieve the specified level are not effective and require alteration. Those which exhibit the specified level may be judged by management to be satisfactory as they are currently taught.

In addition to providing information concerning the effectiveness of training, any good evaluation scheme must also provide information concerning the number, location, and nature of student errors. If it is found that some course, or some aspect of a course, is not effective (i.e., is not meeting specific goals, or subgoals), then it is necessary to identify specific deficiencies so that they can be properly corrected. As noted previously, higher levels of analysis (evaluation) would probably be required to identify the specific nature of the changes that would correct course deficiencies.

Maintaining the Goal Achievement Orientation. For TEA, it is important that the evaluator maintain a clear and consistent focus on what he is trying to do; i.e., determine how well the intended goals of the course are being met. It is recognized that courses may have effects on students other than those that were intended. It has been suspected, for example, that "A" School experience may interfere with retention of behaviors learned in recruit training. This certainly is not desirable, and more definitive information would be needed to conclude that there is a real concern. Investigation of unintended effects is worthy in its own right. It may be desirable to identify and isolate unintended effects so that appropriate action (note that some unintended side effects might be favorable and we would not want to eliminate them) could be taken. Normally, however, an assessment of such effects should not be undertaken in conjunction with an assessment of training effectiveness. Here, the interest should be in determining whether the course has met its goals. Whatever else it may have produced is irrelevant to determining training

effectiveness. Similarly, attempting to discover "what the instructor is really teaching" (i.e., his true goals), by whatever evaluative means, is irrelevant to the issue of whether or not the course goals are being met. If they are not being met, then the instructor's efforts should be redirected toward these ends.

Timing for Assessment. To determine if course goals are being met, it is necessary to be able to attribute measured skills and knowledges to prior training experience. Thus, the time and events intervening between the conclusion of training and the measurement of student achievement must be taken into account. Current attempts to assess the effectiveness of training typically occur after the course graduate has been on the job for some specified period of time (often 6 months). During this interim, it is highly probable that the individual will have undergone on-the-job training (OJT), experienced various types and levels of involvement in various facets of job operations, observed the performance of other individuals, and also discussed various job functions. Given these considerations, the results of any untimely evaluation cannot be interpreted unequivocally. Some aspects of the individual's performance may truly be due to the formal training he received. But others are also undoubtedly due to his experiences between training and assessment. Thus, what we can conclude about the previous training is limited. To avoid this problem, the best time to measure trainees to determine course effectiveness is usually immediately upon completion of the course.

OBTAINING EVALUATION DATA. Once the information needed for evaluation has been identified, it is necessary to select an approach for obtaining it. There are a substantial number of techniques that could be used for obtaining data of interest. Selection from among the alternative approaches should consider factors relevant to the given situation. Ideally, the training effectiveness of a course should be assessed immediately upon completion of that course by appropriate testing routines. Since the objective is to assess training effectiveness, and validity is a special (but related) issue, it is believed that an end-of-course measurement is the most direct and relatively unconfounded measure of training effectiveness. To the extent that this cannot be accomplished, it is necessary to obtain data from the operational setting. The following section describes a number of means for data collection and discusses relevant considerations for their selection and use. Before proceeding to that section, however, evaluation plans under which TEA can be accomplished systematically and objectively are briefly discussed.

PREPARATION OF EVALUATION PLANS

A detailed evaluation plan should be developed, reviewed, and approved in advance of TEA of a course. Once prepared, this plan should be used to guide the conduct of the TEA. Adherence to the plan will

assure the reasonable conduct of the TEA and, consequently, the production of meaningful and usable results.

The basic elements that should be contained in evaluation plans are listed below. A detailed discussion of each of the procedural steps is not given since they largely represent summary statements of material previously presented. No attempt is made to suggest specific responsibilities for preparation, review, and approval of evaluation plans. The intent, rather, is only to delineate the areas that a skilled evaluator would address in preparing to assess particular courses.

At a technical level, evaluation plans should address and provide details on the following minimum items:

- The goals to be met by the evaluation (e.g., to determine if the course is effective and in what areas specific strengths and weaknesses exist, or to determine the effects of recent course revisions)
- A description of the course and its goals
- The data to be collected on students to reflect goal achievement
- The standards (or criteria) to be employed to determine the acceptability of student performance (both individually and/or as a group)
- The techniques that will be used for collecting the data (e.g., questionnaires, interviews, performance testing)
- The details of how the measuring instruments will be selected, modified, or as appropriate, developed
- The details of how the data will be processed and what summary measures will be used to reflect course effectiveness (e.g., measures of central tendency, performance of critical tasks, number of objectives achieved, percentages, percentiles)
- The schedule for data collection and completion of the course evaluation.

Procedures must also be established for reporting the results of the TEA after it has been accomplished. The findings should address the information need for which the evaluation was conducted. Deficiencies should be noted and necessary corrective action identified. At management levels, necessary actions should be taken to insure that appropriate changes are made. Subsequently, the whole process should be repeated to determine if the changes have resulted in a more effective course.

Management levels should review and critically evaluate all elements of proposed plans to insure their adequacy. Training management must also provide the necessary support and logistics for the evaluation. Plans written at higher command levels should probably address the logistics of evaluations; i.e., schedules (e.g., number and identity of courses to be evaluated) and number of graduates to be "assessed" from each course. (It is not necessary to measure all graduates to determine the effectiveness of a course. A sample will provide sufficient information.)

SECTION IV

EVALUATION DATA GATHERING OPTIONS

The method or approach used to acquire information about graduate performance is a critical element in the evaluation process. Competent decisions regarding the value of training and the desirability of changes to it require the availability of reliable and valid data about training effects. There are a substantial number of methods that may be used to obtain the necessary information. At present, the military tends to rely almost exclusively on the use of various types of questionnaires for obtaining such data. But, while questionnaires may be appropriate in some cases, other methods may be better suited and produce more meaningful and definitive data in others. The choice of method should not be arbitrary, but should be based on elements of specific evaluation situations.

In this section, various methodological options which are suitable for use in TEA are discussed. The discussions presented are intended to familiarize training managers and potential evaluators with a range of evaluation data gathering options (EDGO). This increased familiarity should assist those responsible for evaluation to select appropriate methods. Also, various attributes of these options are identified and discussed as they affect the rational choice of methodology for obtaining evaluation data. The discussions are also intended to help training management realize that choices among options can be made intelligently but technical assistance is desirable regardless of which options are chosen.

ATTRIBUTES OF OPTIONS

The selection of an "appropriate" means for obtaining accurate trainee performance data should be based upon desirable characteristics, or attributes of the available options. Options differ in the degree to which they may afford reliable and valid data about training effects. They also differ in ease of use, costs of employment, and in various other ways. These attributes have important implications for option selection. The most significant attributes to be considered in the decision to use a particular option are discussed below. They may be classified as either technical or practical.

TECHNICAL ATTRIBUTES. The technical attributes of an option should be of primary concern in choosing an option for collecting data in a particular situation. They alone determine the value of the TEA data collected and its ultimate usefulness for training improvement. These technical attributes include, most importantly, the aspects of reliability and validity. The "obtrusiveness" of an option may also be considered as a technical attribute of it.

Reliability and Validity. Reliability and validity are indispensable to the production of meaningful and useful evaluation results. If use of an option will not result in the gathering of reliable and valid trainee performance data, it is pointless to conduct a TEA. The results of an evaluation performed from unreliable or invalid data will be meaningless. The results will be uninterpretable, and it will not be possible to attribute them to any prior learning experiences of the students. Note that reliability and validity refer to the data that is (can be) obtained through the use of an option and not to the option itself. For purposes of the discussions here, however, they will be considered as attributes of the options. An essential point that must be remembered--and emphasized--is that options should be selected on the basis of their potential for the production of reliable and valid data. Achieving that potential requires that certain procedures be carefully followed in using the option. Evaluation personnel may require professional assistance to deal with these technically complex matters.

Reliability. In the TEA context, reliability refers to the consistency, or repeatability, of obtained results. If it is reliable, the method followed to obtain data will produce equivalent data each time it is employed. Ideally, procedures will be employed both to check and to enhance method reliability prior to its full scale employment in a TEA. Unless consistent results (data) can be obtained through use of a particular option, there is no way of determining how well students really achieve course goals. Thus, it will not be possible to determine the validity of the obtained data.

Validity. Validity is the single most important attribute to consider in the selection of an option. It refers to the degree to which obtained data reflect what they are supposed to reflect. In TEA, valid data are those which accurately reflect the degree of achievement of course goals. Thus, if a course is intended to produce certain defined skills, then actual measures of trainee skills are intrinsically more valid than scores on a paper and pencil test or opinions obtained from a questionnaire. Again, however, it is the procedures that are followed in using the option that determine the validity of the data obtained by it and not the option itself.

Obtrusiveness of Options. Evaluation data gathering options also differ in obtrusiveness. Obtrusiveness may yield invalid data because the person being observed (or the person providing the data) may be aware that he is in an evaluation situation. Thus, he may react in such a way as to change his responses or judgments from how he would normally respond (see Anderson, et al., 1973). Options which employ objective approaches to gathering data (e.g., achievement proficiency examinations or performance measurements) are much less influenced by the obtrusiveness-reactiveness factor than those which collect subjective data. There is considerable latitude in the interpretation of what is the "correct"

response to give in subjective options (such as questionnaires or interview methods). In more objective assessments of trainee proficiency (i.e., objective scoring of observable performance) there is much less to distort or invalidate the data.

The concept of obtrusiveness, in a second sense, refers to the notion of interference or disruption of normal operating routines. Thus, it is not reactiveness of the respondents but disruption of their work which is of primary concern. Interest has been expressed in the use of nonobtrusive (i.e., noninterfering) data gathering techniques. To the extent they are employed, they are subject to the same considerations of reliability and validity as any other option.

PRACTICAL ATTRIBUTES. The technical attributes of options should be of primary concern in selecting an option. It is recognized, however, that practical limitations may mandate the selection of less desirable options than an evaluator might otherwise want to employ for TEA. Thus, trade offs may be necessary to select an "optimum" technique for a given situation.

The "practical attributes" of options ultimately equate to the various types of costs that may be entailed in using a particular approach for TEA. The use of any option for gathering evaluation data involves costs. Most of the costs incurred arise from personnel requirements (either in numbers of personnel and/or in the required levels of experience). Among the personnel cost attributes of an option are the following: (1) cost to develop the data collection instruments and procedures, (2) cost to administer the data collection instruments (i.e., actually gather the data), and (3) the cost to "score" and analyze the collected data. Nonpersonnel costs; e.g., special testing equipment, supplies, or facilities, are typically minor contributors to the cost of employing an option.

Development Costs. Options for data gathering differ in terms of the time and effort required to prepare for data collection. There may be the costs of establishing, reviewing, or validating training objectives. There may be varying amounts of costs associated with the development and tryout of test items or questionnaire items and the refinement of procedural instructions for use of the data collection instrument. Some options are more costly than others because development of data collection instruments requires considerable personnel skill and experience.

Administration Costs. The data collection process also involves costs which are in addition to the costs of the usual administrative staffing. Options differ in regard to administration cost because of their differing spans of control. Data gathering situations range from "one on one" to "one on many." Obviously, administrative costs are highest when the option dictates that one data gatherer can measure only one trainee at a

time. If there are large numbers of trainees, computers can administer some tests inexpensively. However, this cost advantage may be offset by the development type of cost cited above. Again, personnel expertise is a potential contributor to the cost of employing an option. Finally, there is an "interference" cost. Trainee time will be required for the collection of data--time which in the eyes of many might have been spent more profitably doing something else.

Scoring and Analysis Costs. A third type of cost associated with an option concerns the scoring of trainee behaviors and analysis of the data obtained via the option. Options may differ considerably in these costs depending upon the amount of effort and personnel sophistication required.

In this context, the cost of scoring the data is directly related to the development or availability of standards for determining the acceptability of trainee performance. In some cases, standards for judging trainee performance may be readily available. Hence, costs will be negligible. But, in other cases, it will be necessary to derive them from relevant sources. Standards may be dictated by; e.g., equipment tolerances, documented operational requirements, "usual practices," or tactics. Expert opinions regarding satisfactory performance may also identify standards. Sometimes, the evaluator may only be able to derive performance standards by systematic observation of individuals performing a particular skill. In such cases, the development of standards for trainee performance may incur substantial personnel costs.

Personnel costs also accrue from data analysis considerations arising out of employment of the various options. The interpretation and use of evaluation data can require considerable statistical sophistication. Interpretation of gain scores derived from pretests-posttests of trainees and other forms of trend analysis requires consideration by skilled, experienced psychometricians or statisticians. Opinion data (i.e., ratings of supervisors or responses on questionnaires) concerning training effectiveness also require close scrutiny by knowledgeable data analysts.

OPTIONS

Various options that can be used to gather evaluation data are described next. Pertinent attributes of each option are discussed to assist evaluators in selecting an option for particular situations.

DESCRIPTION OF THE OPTIONS. Options can be categorized on the basis of whether they acquire objective or subjective data and their appropriateness for the assessment of skill or knowledge. They can also be described in terms of their technical attributes and cost. Unfortunately, the situations in which training effectiveness must be determined vary so

widely that a singular characterization of an option for universal application is not possible. A study of Navy training is needed to determine those situations in which there are a sufficient number of common elements to warrant the prescription of a specific data gathering procedure. In the interim, a number of options are described here in terms of their amenability to the assessment of reliability, validity, obtrusiveness, and cost. It should be obvious that this information combined with task-specific statements of objectives and task/training analyses will dictate the ultimate utility of each data gathering option. Subsequent to these discussions, a logic for the selection of a specific option which takes both technical and situational considerations into account is presented.

Objective Data Gathering Procedures. Objective data gathering procedures, subsumed under the heading of proficiency or achievement tests in the psychometric literature, are the best means of measuring the effectiveness of training (Cronbach, 1960, p. 361). Objective proficiency tests share a common characteristic--the response to the tasks they set would not be possible or accurate without the benefit of training. Thus, such procedures provide a measure of training effectiveness. If applied in a programmatic way, they will provide a fair basis for grading trainees and indicating levels of achievement.

While this discussion is concerned primarily with the assessment of training effectiveness, it is interesting to note that training effectiveness data could be a "fringe benefit" of programmatic proficiency testing. Such a testing program could be justified on the accrual of training management benefits (Stuit, 1947) as well as its training effectiveness assessment capability.

Paper and Pencil Achievement Tests. The most familiar and commonly used of the objective EDGOs is the classic paper and pencil test with multiple choice item format. Such tests are excellent for acquiring data on achieved level of job knowledge. In addition to providing test data on proficiency at recalling or recognizing correct facts or principles, this option may also be used to determine knowledge of procedures, symbols, and the application of knowledge. These latter aspects of proficiency testing with paper and pencil are less well known to the training community.

The item format of a paper and pencil test may be designed to test actual job performance which would normally be considered "behavioral." For example, a test task which asks the testee to refer to a picture or diagram of a piece of equipment and point to the location(s) for applying test equipment probes correlates well with the task performed in a work environment (see "TAB Tests" in Glaser, et al., 1954).

Paper and pencil proficiency tests have several virtues: many items can be administered to a group of trainees in a relatively short amount of time, conditions of administration can be standardized, specific training content problems can be diagnosed, and reliability and validity can be determined in a relatively straightforward manner. The problems arising from the employment of this option primarily concern the expertise of the personnel who develop, administer, score and analyze the data, or make evaluation judgments. Good paper and pencil tests cannot be designed and used effectively without a working knowledge of the field of testing.

Factors which affect reliability and therefore require technical knowledge include judgments regarding test length, sampling of item content, item format, item difficulty, and characteristics of the test population. Such expertise is also required to select and make judicious use of the specific statistical methods used to estimate reliability and validity. Given that the reliability of a measure can be established, its validity may still be indeterminate. Probably the most important issue in regard to establishing the validity of a measure is establishing the criterion against which to assess training effectiveness. Conceptually, this task is deceptively simple. Operationally, it is extraordinarily difficult. The basic problem is to determine what it is that we wish to measure. For example, we may wish to assess the rate at which a trainee acquires information, the specific level of knowledge attained, his ability to perform in subsequent courses, or his ability to perform in the Fleet environment. Each alternative requires a concise statement of training objectives which are based on behavioral objectives as determined in a task/training analysis. In each of these instances an additional requirement is imposed by the need to define successful performance. Also, the extent to which the performance has been confounded by such factors as previous experience, motivational variables, educational and/or organizational environment must be assessed.

Clearly, allowing an untrained person to devise and use his own test provides little information in evaluating a course (Byars and Crane, 1969). Unfortunately, the typical Navy instructor needs a great deal more training in this highly crucial area than is presently received at instructor training school.

The cost attributes of paper and pencil testing can be hinted at only in the most general sense because of the diversity of specific course evaluation requirements. Development costs are fairly high because of the requirement for professional assistance or expertise to do an acceptable job. Administration costs are quite low. Development of scoring and analysis procedures for the paper and pencil option are moderate, again requiring professional assistance. In general, these cost attributes can be offset when the course to be evaluated has an annual man-hours trained figure that is high; i.e., long course duration and/or large throughput of trainees. The cost attributes may even be

"paid back" or amortized if the paper and pencil proficiency test is incorporated into a continuous course quality control system.

Work Sample Proficiency Tests. There are two options which consist of systematic observation and scoring of the performance of trainees in a test situation which reproduces a significant sampling of actual job operations. The two options differ mainly in the degree to which the work sample is a simulation of the job in a real job environment. In one instance measurement takes place in the actual job environment; in the other, measurement takes place in a simulated work environment. Practicality, safety, standardization of administration, and cost determine the desirability of choosing between these two options. For example, if training devices are used that can effectively simulate the required features of the job and commensurate job skill, then the choice should obviously favor work sample proficiency assessment in a simulated environment. For most courses in the school environment, work sample proficiency testing would employ a simulated work environment.

The work sample proficiency test is the most appropriate for evaluating the effectiveness of technical skill training. They are most useful for determining what a trainee or course graduate can do. They must be employed in evaluating the skills that cannot be assessed via a paper and pencil test; e.g., psychomotor tracking tasks, soldering, typing. This kind of testing is best conducted in a controlled setting. However, with a well constructed observer checklist, they can be conducted in on-the-job situations. Well developed work sample proficiency tests can acquire reliable data that are both valid and highly diagnostic of training problem areas.

The objectivity and reliability of the data gathered in this manner depend upon the same type of technical considerations presented in the discussion of paper and pencil tests. Consequently, professional assistance is a necessity in the development of the testing procedures and the development of the performance scoring criteria. Wherever possible, the observer/scorer should use speed and/or accuracy criteria so as to avoid ambiguities of scoring since individual judgments of performance quality are notoriously unreliable and should be used carefully and sparingly, or not at all. What should be scored in this type of testing situation are the performances of a process and/or the product of that process. Products are most easily scored in terms of the degree to which they meet specifiable standards. To score processes or procedural finesse, the behavior of the individual being tested must be observed and data systematically recorded for subsequent scoring.

Because of scoring sophistication requirements and the need to actually observe the performance of the individual being tested, work sample proficiency tests cannot be administered to large groups simultaneously. This administrative limitation dictates that both tasks and trainees must be sampled. Such sampling requires professional assistance

in designing the sampling procedures that will provide the greatest scope of inference at the least cost.

Work sample proficiency tests are the most costly of all the available options because of the need for highly competent professionals to develop the procedures and materials and to train observers/administrators. The cost to develop scoring and analysis procedures is high for the same reason and because of the necessity to acquire data on work samples and performance criteria from subject-matter experts. However, the initial cost of developing work sample proficiency tests can be amortized and paid back in the same manner as paper and pencil achievement tests in longer duration, higher throughput courses. Finally, simulated work environments may increase measurement costs if additional equipment and separate facilities are required.

Subjective Data Gathering Procedures. Subjective data gathering procedures are the most widely employed in current practice. Compared to the objective data gathering procedures, subjective procedures are less expensive to employ.

Unfortunately, the subjective data obtained via these procedures are opinions about the proficiency of the trainee/graduate, rather than direct measurements of skill or knowledge. That is, the criteria used by those whose opinions are solicited may be unknown and idiosyncratic. There are also a number of other technical and practical considerations which affect the usefulness of data collected by subjective means.

Questionnaires. Questionnaires exist in many forms. The basic format consists of a group of printed questions designed to elicit opinions from trainees, course graduates, instructors, or job supervisors. The items on the questionnaire may take the form of open-ended or fixed-choice questions. In the former, the respondent answers in his own words while in the latter he is required to select a choice from provided alternatives. Questionnaires may also contain checklists or rating scales. Obviously, they provide a high degree of flexibility and may be used to ask questions concerning a wide range of topics. However, this flexibility has associated with it a number of technical considerations which require appropriate expertise. In making up the questionnaire, the evaluator needs to select the appropriate form of his questions to obtain the kind of opinion he desires. Add to the foregoing the fact that all the requirements for assessing the reliability and validity discussed earlier are also applicable here.

There are some additional measurement related problems which are magnified when questionnaires are employed. First, there are problems associated with the validity of the responses. For example, respondents may not be knowledgeable enough to respond to questions. If they are knowledgeable they may not be able to communicate accurately enough to

provide useful data. Even if they are knowledgeable and have good communication skill, they may not be motivated. Another source of difficulty arises from the fact that in the process of identifying problem areas, one may, in effect, be placing one's self or superiors on report. Hence, the problem of acquiescence to the "proper" response and in some cases dishonesty of responses may be a source of data invalidity.

Not to be ignored are the more practical issues associated with the administration of questionnaires such as mailout return rates (notoriously poor), long lead times for sending out and recovering questionnaire data, or gaining access to the study population. While these appear simple to resolve, they become inordinately burdensome when multiplied by the various levels of command.

All of the military services use questionnaires of various descriptions for obtaining opinion data about the performance (or knowledges) of individuals who have received training. The value of questionnaire data for training improvement is largely unresolved since they do represent opinions rather than actual, factual information about ability. Questionnaires do not directly provide measures of operational performance. A recent TAEG study (Dyer, Ryan, and Mew, 1975; Dyer, Mew, and Ryan, 1975) has indicated that well-designed questionnaires, based on structured job performance requirements, can be used to obtain information to identify training problems. Care must be taken in interpreting the data obtained via this technique, however, since they do not directly constitute measures of training effectiveness.

The advantages of using questionnaires for the evaluation of training are that they are relatively inexpensive and easy to administer in the sense that they do not require a one-on-one evaluation situation. Opinion information relevant to both training and job performance can be obtained. The use of the questionnaire obviously is not the same as a standardized testing situation, and respondents may vary considerably in the amount of time and care they take in answering questions. Despite the great popularity of the questionnaire method for obtaining evaluation data, it has serious problems associated with its use. At best it only provides subjective opinions, not objective performance data on which to evaluate training.

Interviews. An interview may be considered as an oral questionnaire. It involves a conversation between an interviewer and a respondent which permits the interviewer to obtain information about a person or his performance. In evaluating training programs, the interview may be used as a means of obtaining data on trainee background variables (e.g., family, education, interests, attitudes) and on student opinions about the training program, its materials, and the instructor. However, much of this background material is easily obtainable from the trainees' personnel records and, therefore, should not be obtained in an interview

format. In addition, interview data from instructors and other personnel (e.g., job supervisors and commanders) can be used to obtain recommendations for the content and conduct of the training program.

In order to obtain accurate interview data, the evaluator needs to obtain a representative sample of trainees, supervisors, and/or their commanders. Also, it is necessary to properly train the interviewers to avoid bias and insure consistency of technique. The interview is particularly prone to low levels of interrater reliability.

Much like paper and pencil tests, items in the interview must be assessed for reliability and validity. But, the interview is unique in that the reliability of the interviewer must be assessed in addition to the interview questions. Thus, considerations of the interviewer's personal characteristics, his interpersonal skills, and his identification with the sample population become relevant issues. It must also be remembered that the interview is conducted in a social context and is therefore affected by the reactive nature of the interpersonal situation. Only careful attention to construction of interview questions and training of interviewers can minimize the problems associated with this option.

The advantages of using an interview for evaluation are (1) it can be used to evaluate complex training, (2) the interviewer can check on the information he is given, and (3) it is adaptable to nearly all evaluation situations. It is also costly to use for evaluation purposes. But, sometimes the interview is the only technique available to obtain certain evaluation information.

Records and Reports. Records and reports may provide either objective or subjective information for TEA. If students from a given training course consistently perform poorly on the job as indicated by records and reports of job performance, this strongly indicates a problem with the training program. In those cases where trainees are sent to the job from more than one training source, records and reports can provide a direct comparison of job performance which would reflect, at least to some extent, the relative effectiveness of the two sources of training. Records and reports might include job diaries (if available), superior's logs, absenteeism and tardiness on the job, work production on the job, time in grade, letters from commanding officers and job supervisors as to the quality of training shown on the job, time to complete the training course, attrition from the training course, frequency of setbacks in the training course, attendance in the training course, and scores from the training course. To be of value, records and reports must be accurate.

The advantages of using records and reports for training evaluation are that they are simple and easily obtained, inexpensive, relatively easy to administer, and amenable to the evaluation of routine jobs.

However, many records and reports are so irrelevant or ambiguous that they are not useful. Costs of these techniques vary widely and, beyond administration, can be largely attributed to time lost from the job.

Variations of EDGO. Each of the standard options for evaluation listed above can be used in conjunction with "special" techniques. Special techniques include criterion-referenced measurement (CRM), computer-managed testing (CMT), time-series analysis, pretest/posttest design, and secondary analysis. Criterion-referenced measurement is basically a validating technique that can be applied to evaluation options. Theoretically, it can be applied to any option, but, practically, it is best applied to achievement tests and tests of on-the-job tasks; i.e., objective options. It would probably be difficult and impractical to apply the CRM to subjective options and records and reports.

Computer-managed testing provides a tool or a method for presenting and scoring an evaluation test. Theoretically, it too could be used to supplement the six basic options. It can probably be best used with the achievement test and the questionnaire and least well with a performance test or an interview. Time-series analysis is basically an analysis technique under which the data obtained from the basic evaluation options are plotted over time to show trends in evaluation results. Pretests and posttests involve an experimental design or strategy for testing for training effects. They can be used with any of the six evaluation options described below but have been used most frequently with achievement tests. Secondary evaluation consists basically of a review and reanalysis of data derived from one of the six basic evaluation options. Secondary evaluation can be applied to the data derived from any of the six evaluation options.

EDGO SELECTION. The characteristics of the training to be evaluated and the purpose of the evaluation interact with the attributes of options to determine which option or mix of options to employ in particular situations. A selection logic is presented to assist the potential evaluator in making systematic judgments with regard to the use of the various options. Following this logic will assure that relevant decision trade offs will be made in arriving at a desirable evaluation program. It must be remembered that technical expertise may be required to use selected options properly.

Logical Considerations. Figure 1 formalizes the EDGO selection process and summarizes the logic for the selection of options. It is based on the assumption that there is a need for an evaluation program. Some features of this decision analysis are worth noting. Decision trade offs are addressed in a sequential order and are identified by diamond shapes. If existing data will satisfy evaluation needs (A) no further action is necessary to select options. However, if additional data are required, the second decision (B) has the greatest impact on option

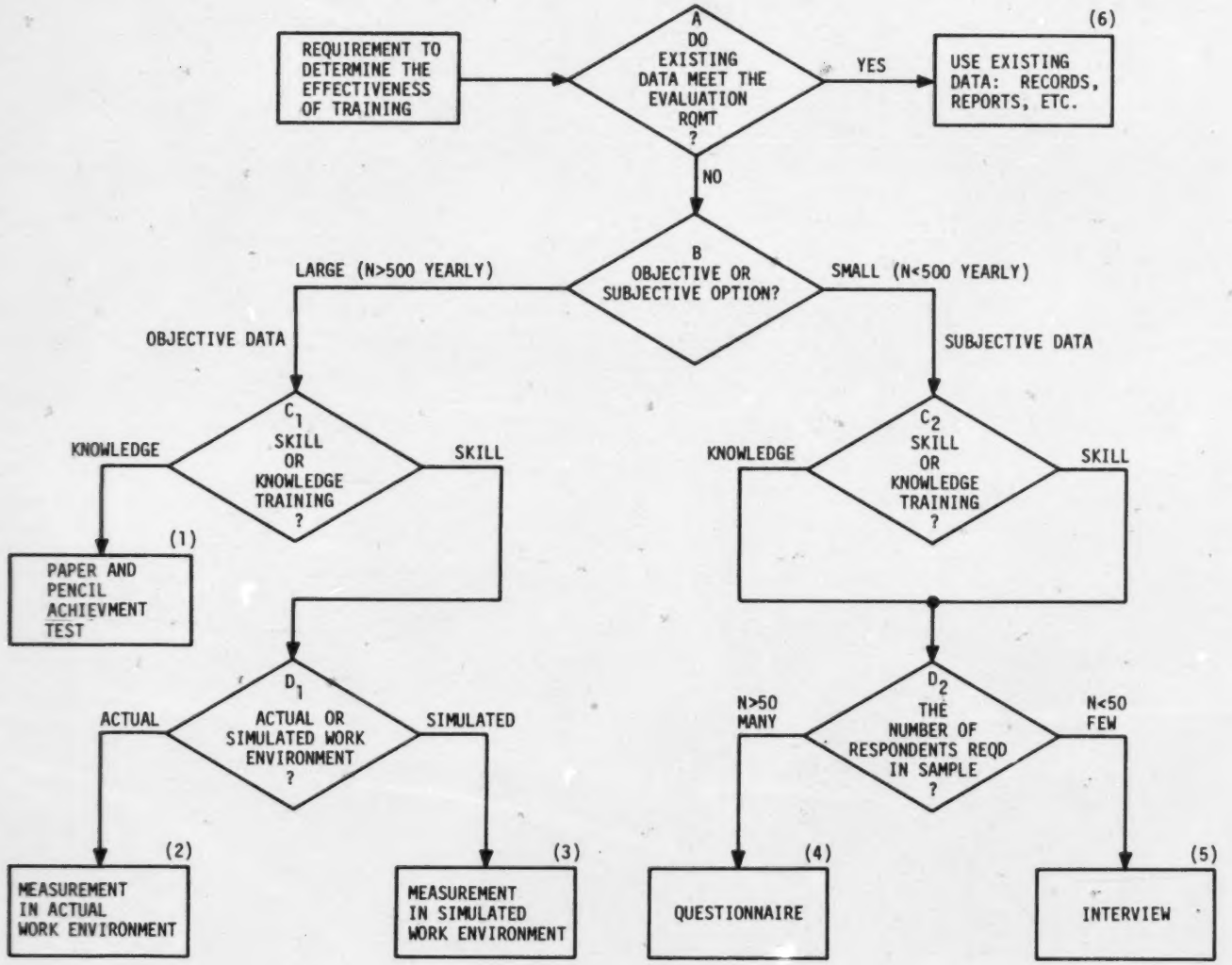


Figure 1. EDGO Selection Rationale

selection. It leads directly to the choice of objective proficiency measurement procedures or to procedures that acquire opinions. If a course to be evaluated contains both skill and knowledge components, then two or more options may be indicated by the selection logic (C_1). If the evaluation data may consist of opinions, then the skill and knowledge distinction has no effect on options selected (C_2). Skill testing, in the case of objective data (D_1), and the number of respondents in the case of subjective data (D_2) require an additional decision with regard to specific options.

Throughput or man-hours trained is a significant consideration for the option selection process. High volume courses are of utmost concern in assessing the effectiveness of training. The higher the personnel throughput, the greater the desirability for objective, reliable, valid, and diagnostic data. Impact of throughput is greatest in the trade off decision that determines whether objective or subjective data will be required (B). Finally, the six (numbered) terminal boxes in figure 1 contain the names of the recommended generic options for the gathering of training evaluation data. The rationale for each decision trade off point is discussed below in more detail to provide a firm understanding of the entire logic sequence.

A. Do existing data meet the evaluation requirements? Existing data should always be examined first to determine if information contained in records or reports (6) can be used to indicate the effectiveness of training. Such information may also be diagnostic of training problems. Records or reports having the most utility for evaluation are those which result from a well organized quality control system which routinely produces data for evaluation and diagnosis of training problems. Such quality control systems for Navy training, though sorely needed, do not presently exist. However, failure to take note of all available and relevant data could result in the unnecessary use of some other EDGO.

B. Should an objective or subjective EDGO be employed? The answer to this question bears directly on the issue of whether objective or subjective data best satisfy evaluation requirements. Must the evaluator know what a trainee/graduate can do as a result of training, or can the evaluator be satisfied with what someone (supervisors, peers, or trainee/graduate) thinks the trainee/graduate can do? In general, objective data are always preferred. When they cannot be obtained, it may be necessary to use options that gather as data the attitudes, perceptions, and opinions of others.

A number of factors favor objective data over subjective data for training evaluation use. Some prime considerations are inherent in what might be called variables of investment and concern for throughput efficiency or output quality. Investment refers to the magnitude of resources devoted to a course and includes such items as personnel

costs, facilities and equipment costs, and training material development costs. The rationale for this criterion is simply that the greater the resource investment in training, the greater the need for determining fairly and objectively the effectiveness of the training.

When there is "concern" that the output of a training course may not be adequate, objective data are needed to define the nature of the problem. If there is "no concern" for a course, then subjective data may be routinely employed to monitor for potential problems. When a possible problem is identified, it is necessary to determine the value of and direction of further efforts to gather objective and diagnostic data.

A critical value of investment or concern on which to base an option judgment can only be arbitrarily stated. For example, it appears reasonable to assume that courses with annual throughput in excess of 500 trainees/graduates represent the "right" level of investment to justify selection of objective data gathering options. And, when "concern" that something may be wrong with a course reaches the subjective level of 50-50 odds, then the decision to acquire objective data also appears reasonable. Again, it must be stated that these are only suggested levels of the critical values that may logically support and justify a decision to select objective versus subjective data options. The actual evaluation context also affects the decision as to whether opinions or subjective impressions about training are acceptable data or whether objective, factual data are required.

C. Is the type of learning, that makes up the training to be evaluated, the acquisition of skill or knowledge? Learning results in either new knowledge or new skill, or both. Knowledge is essentially information; skill is behavior that improves with practice. Most training consists of both skill and knowledge acquisition by the trainee. However, it is important to note the difference when selecting an EDGO. The data gathering option employed in the evaluation of the "skill" part of a training course may be inappropriate for evaluation of the "knowledge" part of the course. The measurement of "knowledge" as the evaluation goal using objective data should result in the use of paper and pencil achievement tests (1). However, the objective measurement of skill and subjective assessment of skill or knowledge requires an additional decision.

D. Factors affecting the selection of a specific EDGO. Different factors affect the selection of a specific option given the decision to acquire either objective or subjective data. In the case of the former, it must be determined if measurement should take place in the actual or simulated environment (D_1). For subjective data, the choice of an option depends greatly upon the required sample size (D_2). Option choice, of course, should be moderated by consideration of the technical and practical factors identified earlier.

Examples of EDGO Selection. Examples are presented here to illustrate how the EDGO selection logic can be applied starting with an evaluation requirement and ending with one or more options selected (see figure 1). Two extreme cases that an evaluator might encounter in Navy training evaluation are hypothesized. These scenarios have been deliberately contrived to be quite disparate to show how different training evaluation situations indicate the use of different EDGOs for obtaining appropriate assessment data.

Case A. The requirement to evaluate this course is in response to a flurry of complaints from the Fleet, over a period of 2 months, generally about the incompetence of recent graduates. Examination of existing records reveals two suspicious trends: the number of minimum aptitude waivers has increased for the past year, but starting 6 months ago, the academic attrition has become low and steady at 4.9 percent. Since the setback rate for the course has not changed appreciably in 2 years, it would appear that standards have been lowered. However, there are no records of test performance by previous classes. This, coupled with the fact that the annual throughput of this course is just over 4000 graduates, makes a compelling justification for seeking objective data about the effectiveness of this investment.

This leads to the next decision point which depends on the skill and knowledge portions of the course (C_1). Since this course trains both in about equal amounts, two options to gather evaluation data are identified. The decision is easily reached to assess knowledge training via a paper and pencil achievement test. The skill portion of the training should be assessed using data acquired through the administration of a performance-based proficiency test.

Again, a decision must be made as to whether or not the performance of trainees/graduates will be measured in an actual or simulated work environment (D_1). Since the real job tasks for which the training was designed are expensive and potentially dangerous to perform, the evaluator should select option 3, a simulated work sample. If the job tasks were relatively inexpensive and safely performed, then option 2 would have been selected, unless other factors ruled it out.

Case B. This case is much less difficult than case A. The impetus for evaluating this course is simply a routine requirement that all courses of its type should undergo some form of evaluation every 3 years. There is no reason to be concerned about it; i.e., no complaints. The course has a relatively small throughput of about 100 graduates per year. The content of the course is predominantly skill training--90 percent skill and 10 percent knowledge training.

No previously accumulated records or reports are available. The course has a relatively small throughput. The logic of EDGO selection in figure 1 suggests that subjective data would be most economical to

acquire. Data which are not as objective and diagnostic will be gathered, but they are deemed acceptable.

Figure 1 shows that the skill or knowledge portions of the course have no bearing on option selection for subjective data. Actually, all that remains to be determined is whether to employ option 4 or option 5, or both. Cost is an important consideration. If the number of respondents needed were large, say 50 or more trainees/graduates, then the least expensive subjective option, the questionnaire (4), should be selected. Or, if the number of respondents required is small, say fewer than 50, then the more expensive subjective option, option 5, the interview (structured or not), should be chosen. In this hypothetical case, since interviewees were both representative and easily reached, the evaluator chose to interview a small sample of graduates and their supervisors.

Thus, the kind of EDGO selected is a function of a few key decisions based on relatively simple, but critical, information requirements. It is hoped that this selection logic will guide the potential Navy training evaluator to appropriate options for evaluating courses. Key decisions regarding option selection must be based on sound trade offs between the practical and technical attributes of each option and the characteristics of each specific evaluation situation.

INTERPRETATION AND USE OF EVALUATION DATA

Competent use of any of the evaluation data gathering options described above will result in the production of data about the degree to which student performance reflects the achievement of course goals. The reliability and validity of the obtained data will differ as a function of the method used and procedures followed in gathering the data. Thus, the utility of the data for meeting specific evaluation purposes will vary. For purposes of this discussion, however, assume that data reliability and validity are acceptable and that the obtained data "contains" correct information about the student behavior brought about as a result of training. The next problem is to interpret the meaning of these data and correctly use them for intended purposes.

All too often the mistake is made that simply obtaining data is sufficient for many training design or evaluation problems. For example, many have verbalized that the solution to training problems is to perform a task analysis of the job for which training is (to be) given. Thus, many task analyses are done which do not achieve the intended training analysis purpose. The resulting data are either not used at all or not well used. Someone must interpret the obtained data and decide how it may best be used in training design. This requires a translation of task information into training objectives, allocation of these objectives to various portions of training, and development of appropriate content and methods for training. This analysis problem is complex. Similarly, in the TEA domain, there seems to be a belief that possession of a means for obtaining data about trainee performance is sufficient to insure effective training.

Current CNET instructions regarding training appraisal (e.g., CNET Instruction 1540.3A) underscore the need for training organizations to obtain data reflecting the post training competency of their students. However, little guidance is provided for using such data for training improvement after they have been obtained.

The uses to which performance data can be put are limited by the data themselves. However, to the extent that the data collection was accomplished in accordance with some predefined purpose (see section III), the form of the data should be acceptable. As noted in section III, the purpose of TEA may simply be to determine the current level of effectiveness of a course. Thus, the obtained data must be (appropriately) summarized. This can be done in a number of ways. For example, course effectiveness could be expressed in terms of percentage of required job tasks that individuals can accomplish, numbers or percentages of individuals who satisfactorily perform job tasks, mean or median scores, etc. While summary scores are acceptable for simply expressing training effectiveness, more detailed information is needed for training course improvement. For this purpose it will be necessary to examine the specific "errors" made by the students on objective tests of their training acquired abilities. (Subjective opinion data is not conducive to error analysis.) The interpretation of test data requires considerable skill.

Suppose the TEA reveals that an unacceptable number of students make the same types of errors on a particular task. The reasons for the errors could be due to training deficiences (e.g., not enough time devoted to its training, lack of hands-on practice during training, poor instructor presentation of the task.) However, the poor performance could be due to factors other than training. Conceivably, the task, as it is structured, could itself be intrinsically too difficult for its successful (or easy) performance. In this case, changes to the task structure might be required. That is, operations might need to be changed--not training. Another possible reason for unsatisfactory performance of a task lies in the students' (or graduates') abilities. In the example cited, poor task performance might be because students trained in the course do not have the abilities (or aptitudes) required for successful performance in the operational situation. Rather than changing the course, it may be necessary to change student entry level qualifications. Thus, considerable expertise is required to interpret TEA data correctly.

If it is concluded that observed deficiencies are probably due to training inadequacies and that they can be corrected/alleviated by training, then the data must be appropriately used to alter the training course. What specifically to change in training requires further analysis (see section III) and study of student errors. Deficiencies could be due to inappropriate content or instructional strategies, lack of training equipment, etc. This process of attributing deficiencies to particular aspects of the prior training can also be very demanding and require considerable knowledge on the part of the evaluator. An additional problem is inherent in deciding how to change training to achieve better outcomes. Again, considerable technical and subject-matter expertise will be required to select or develop, for example, more effective instructional strategies.

In summary, the course evaluator's most difficult task begins after data reflecting student achievement have been collected. Those data must be analyzed and the meaning and implications for course changes correctly ascertained. Specific ways of changing the course to correct deficiencies must also be identified or developed, implemented, and evaluated.

SECTION V

CONCLUSIONS AND RECOMMENDATIONS

Conclusions about training evaluation within the Navy are presented here. Recommendations for improving the value of evaluation programs are also presented.

CONCLUSIONS

The full potential of evaluation programs for controlling the quality of training is not being realized within the Navy. There are a number of interrelated reasons for this situation. They include:

- Unfavorable attitudes toward evaluation
- Lack of command emphasis on routine evaluation of training programs
- Unclear assignment of responsibilities for training evaluation
- Inadequate command support and surveillance
- Inadequate numbers of personnel for conducting evaluation programs
- Lack of relevant training for those given evaluation responsibilities
- Lack of independence of training and evaluation functions
- Lack of time and other resources for conducting evaluation, and
- A general lack of technical expertise in evaluation concepts and methodology.

Although this list could be extended, the essential point is that evaluation of training has not been given the attention and resources which are required for maintenance of a high-quality training system. Information is not routinely available about baseline training effectiveness. Such information is needed to determine the value of current training (courses) and to identify areas where improvements may be desirable. Present and planned procedures for obtaining and using training effectiveness information are not optimum. They may fail to yield the information needed for informed decision making about training.

The identification, collection, interpretation, and use of training effectiveness information require considerable technical expertise. This expertise is not currently possessed by individuals assigned evaluation tasks. The trend toward the exclusive use of questionnaires for obtaining data about training effectiveness reflects a lack of familiarity with other techniques that may be better suited for obtaining effectiveness information. A substantial number of methodological options are available for obtaining such information. Their selection and use should be based on consideration for specific elements of the evaluation situation.

RECOMMENDATIONS

The achievement and maintenance of high-quality Navy training demands that evaluation be an accepted, integral part of the system. Training evaluation results reflecting the success or failure of courses in meeting their goals should be routinely available to training management to tradeoff against resources required to operate the machinery of instruction. To achieve this end, much concerted effort is needed by the Naval Education and Training Command (NAVEDTRACOM).

As a first step, greater command emphasis should be placed on training evaluation. Firm policy requiring the conduct of training evaluation should be established. Such policy should include a clear delineation of responsibilities for evaluation. It should also identify reporting channels for dissemination of training effectiveness information and establish a corrective action system for insuring proper use of obtained information.

The TAEG also recommends that a strong evaluation function be established within the NAVEDTRACOM. This function (or group) would be specifically charged with responsibility for conducting training effectiveness assessments. It should function and report independently of the training process. In the TAEG view, this group should develop and provide information to CNET for controlling the quality of the training system. More careful study by TAEG, CNTECHTRA, and CNET staff is needed to define the appropriate charter, structure, and manning for such an organization. A data collection capability could be provided by the Fleet Feedback Data Collection Groups (FFDCG). The FFDCG concept is currently being defined and evaluated.

It is further recommended that a Training Effectiveness Assessment Center be permanently established within the TAEG. This Center would assist the Training Command in planning and conducting assessments of training effectiveness (courses and different instructional methods or media). It would function on CNET request either to evaluate specific aspects of training or to assist training units in preparing for and conducting such assessments. Such assistance could take a

number of forms ranging from evaluations of potential training media to assessment of specific training courses. Likely, in the latter case, TAEG's role should be one of preparing a specific evaluation plan for the assessment as outlined in section IV.

REFERENCES

Anderson, Scarvia B., Ball, S., Murphy, R. T., and Rosenthal, Elsa J. Anatomy of Evaluation: Important Concepts and Techniques in Evaluating Education/Training Programs. PR-73-36. August 1973. Office of Naval Research, Washington, DC.

Byars, L. L. and Crane, D. P. "Training by Objectives: a Comprehensive System for Evaluating Training Programs." Training and Development Journal. June 1969. pp. 38-40.

Chief of Naval Education and Training. Appraisal and Improvement of Training. CNET Instruction 1540.3A (Draft, undated). Chief of Naval Education and Training, Pensacola, FL.

Cronbach, L. J. Essentials of Psychological Testing. Second Edition. New York: Harper & Brothers. 1960

Dyer, F. N., Ryan, L. E., and Mew, Dorothy V. A Method for Obtaining Post Formal Training Feedback: Development and Validation. TAEG Report No. 19. May 1975. Training Analysis and Evaluation Group, Orlando, FL 32813.

Dyer, F. N., Mew, Dorothy V., and Ryan, L. E. Procedures for Questionnaire Development and Use in Navy Training Feedback. TAEG Report No. 20. October 1975. Training Analysis and Evaluation Group, Orlando, FL 32813.

Glaser, R., Damrin, D., and Gardner, F. "The Tab Item: A Technique for the Measurement of Proficiency in Diagnostic Problem-solving Tasks." Education and Psychological Measurement. 1954. 14. pp. 283-293.

Popham, W. J. Educational Evaluation. Englewood Cliffs, NJ: Prentice-Hall, Inc. 1975.

Ricketson, D. S., Schulz, R., and Wright, R. H. Review of the CONARC Systems Engineering of Training Program and Its Implementation at the United States Army Aviation School. Consulting Report. April 1970. Human Resources Research Organization, Fort Rucker, AL.

Stuit, D. B. (Ed.) Personnel Research and Test Development in the Bureau of Naval Personnel. Princeton, NJ; Princeton University Press. 1947.

BIBLIOGRAPHY

RECORDS AND REPORTS:

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. Unobtrusive Measures: Non-reactive Research in the Social Sciences. Chicago: Rand McNally. 1966.

ACHIEVEMENT TESTS:

Adkins, Dorothy C., et al. Construction and Analysis of Achievement Tests. 1947. U.S. Government Printing Office, Washington, DC.

Denova, C. "Is This Any Way to Evaluate a Training Activity? You Bet It Is." Personnel Journal, 1968, 4 (7), pp. 488-493.

Department of the Air Force. Principles and Techniques of Instruction. Air Force Manual 50-9. April 1967. Washington, DC.

Ebel, R. L. Measuring Educational Achievement. Englewood Cliffs, NJ: Prentice-Hall. 1965.

PERFORMANCE TESTS (SIMULATED AND ACTUAL):

Boyd, J. L., Jr., and Shimberg, B. Handbook of Performance Testing. A Practical Guide for Test Makers. January 1971. Educational Testing Service, Princeton, NJ.

Harris, J. H., Campbell, R. C., Osborn, W. C., and Boldovici, J. A. Development of a Model Job Performance Test for a Combat Occupational Specialty: Volume I. Test Development. FR-CD(L)-75-6. November 1975. Human Resources Research Organization, Alexandria, VA.

Highland, R. W. A Guide for Use in Performance Testing in Air Force Technical Schools. ASPRL-TM-55-1. 1955. Air Force Personnel Training and Research Center, Lowry AFB, CO.

Osborn, W. C. Developing Performance Tests for Training Evaluation. Professional Paper 3-73. February 1973. Human Resources Research Organization, Alexandria, VA.

Steinemann, J. H. Comparison of Performance on Analogous Simulated and Actual Troubleshooting Tasks. SRM-67-1. July 1966. Naval Personnel Research Activity, San Diego, CA.

BIBLIOGRAPHY (continued)

INTERVIEWS AND QUESTIONNAIRES:

Brigham, F. R. "Some Quantitative Considerations in Questionnaire Design and Analysis." Applied Ergonomics, 1975, 6, pp. 90-96.

Gorden, R. L. Interviewing: Strategy, Techniques and Tactics. Homewood, IL: The Dorsey Press. 1969.

Kahn, R. L. and Cannell, C. F. The Dynamics of Interviewing. New York: John Wiley and Sons. 1957.

Sellitz, C., Jahoda, M., Deutsch, M., and Cook, S. Research Methods in Social Relations. (Revised) New York: Holt. 1960.

Sinclair, M. A. "Questionnaire Design." Applied Ergonomics, 1975, 6, pp. 73-80.

CRITERION-REFERENCED MEASUREMENT:

Foley, John P., Jr. Criterion-referenced Measures of Technical Proficiency in Maintenance Activities. AFHRL TR-75-61. October 1975. Air Force Human Resources Laboratory, Wright-Patterson AFB, OH. Work Unit Numbers 62703F 17101007.

Glaser, R. and Nitko, A. J. "Measurement in Learning and Instruction." In R. L. Thorndike (Ed.) Educational Measurement (2nd ed.) Washington, DC.: American Council on Education. 1971. pp. 625-670.

Popham, W. J. (Ed.) Criterion-referenced Measurement. Englewood Cliffs, NJ: Educational Technology Publishers. 1971.

Swezey, R. W. and Pearlstein, R. B. Guidebook for Developing Criterion-referenced Tests. August 1975. Army Research Institute for the Behavioral and Social Sciences (Contract No. DAHC-19-74-C-0018), Arlington, VA 22209.

COMPUTER ASSISTED/MANAGED TESTING:

Ferguson, R. L. Computer Assisted Criterion-referenced Testing. March 1970. Office of Naval Research (Contract No. Nonr-624(18)), Washington, DC.

Holtzman, W. "The Changing World of Mental Measurement and Its Social Significance." American Psychologist, 1971, 26, pp. 546-553.

BIBLIOGRAPHY (continued)

Suppes, P. "The Uses of Computers in Education." Scientific American. 1966, 215(3), pp. 207-219.

ASSESSING CHANGE (PRETEST/POSTTEST):

Campbell, D. T. and Stanley, J. C. "Experimental and Quasi-experimental Designs for Research on Teaching." In N. L. Gage (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally. 1963.

Carver, R. "Special Problems in Measuring Change with Psychometric Devices." Evaluative Research: Strategies and Methods. 1970. American Institutes for Research, Pittsburgh, PA. pp. 48-66.

Harris, C. (Ed.) Problems in Measuring Change. 1963. University of Wisconsin, Madison, WI.

DISTRIBUTION LIST

Air Force

Headquarters, Air Training Command (XPTD, Dr. D. E. Meyer)
Headquarters, Air Training Command (XPTIA, Mr. Goldman)
Air Force Human Resources Laboratory, Brooks Air Force Base
Air Force Human Resources Laboratory (Library), Lowry Air Force Base
Air Force Office of Scientific Research/AR (Dr. A. R. Fregly)

Army

Commandant, TRADOC (Technical Library)
Army Research Institute (Dr. Ralph R. Canter, 316C; Dr. Edgar Johnson

Coast Guard

Commandant, U.S. Coast Guard Headquarters (G-P-1/62)

Marine Corps

CMC (Code OT)
CGMCDEC (Mr. Greenup)
Director, Marine Corps Institute

Navy

Assistant Secretary of the Navy (R&D) (4E741, Dr. S. Koslov)
CNO (OP-987P7, CAPT Connery; OP-991B, M. Malehorn; OP-987P10, Dr. R. Smith;
OP-987, H. Stone)
ONR (Code 458, 455)
ONRBO Boston (J. Lester)
ONRBO Chicago
ONRBO Pasadena (E. E. Gloye)
CNM (MAT-03424, Mr. A. L. Rubinstein)
CNET (01, 00A, N-5 (5 copies))
CNAVRES (Code 02)
COMNAVSEASYSCOM (03, 047C1, 047C12)
COMNAVAIRSYSCOM (03, 340F)
CNET SUPPORT (00, 01A)
CNTECHTRA (0161, Dr. Kerr (5 copies))
CNATRA (F. Schufletowski)
COMTRALANT
COMTRALANT (Educational Advisor)
COMTRAPAC

DISTRIBUTION LIST (continued)

Navy (continued)

CO NAVPERSRANDCEN (Code 02, Dr. Regan; Dr. Earl Jones; Library)
NAVPERSRANDCEN Liaison (Code 01H)
Superintendent NAVPGSCOL (Code 2124)
Superintendent Naval Academy Annapolis (Chairman, Behavioral Science Dept.)
CO NAVEDTRAPRODEVCEN (AH3)
CO NAVEDTRASUPPCEN NORVA
CO NAVEDTRASUPPCENPAC (5 copies)
CO NAVAEROMEDRSCHLAB (Chief Aviation Psych. Div.)
CO FLTCOMDIRSYSTRACENPAC
CO NAMTRAGRU
CISO, NTTC Corry Station
CO NAVTRAEQUIPCEN (N-21, N-215, N-131 (2 copies), N-2211, N-00AF, N-00M, PM TRADE)
Center for Naval Analyses (2 copies)
U.S. Naval Institute (CDR Bowler)

Other DOD

Military Assistant for Human Resources, OAD (E&LS) ODDR&E (LT COL Henry Taylor)
OASD (I&L) - WR (LT COL Grossel)
Director, Human Resources Office, Defense Advanced Research Projects Agency
(R. Young)
Defense Advanced Research Projects Agency (CTO, Dr. H. F. O'Neil, Jr.)

Non-DOD

Essex Corporation (Dr. John Collins)

Information Exchanges

DDC (12 copies)
DLSIE (James Dowling)
Scientific Technical Information Office, NASA
Executive Editor, Psychological Abstracts, American Psychological Association
ERIC Processing and Reference Facility, Bethesda, MD (2 copies)